Biomedical Informatics Insights

ORIGINAL RESEARCH

Three Hybrid Classifiers for the Detection of Emotions in Suicide Notes

Maria Liakata[1,2,4], Jee-Hyub Kim[1,4], Shyamasree Saha[1], Janna Hastings[1,3] and Dietrich Rebholz-Schuhmann[1]

[1]EMBL-EBI, Cambridge. [2]Aberystwyth University, UK. [3]University of Geneva, Switzerland. [4]Joint first authors.
Corresponding author email: liakata@ebi.ac.uk, jhkim@ebi.ac.uk

Abstract: We describe our approach for creating a system able to detect emotions in suicide notes. Motivated by the sparse and imbalanced data as well as the complex annotation scheme, we have considered three hybrid approaches for distinguishing between the different categories. Each of the three approaches combines machine learning with manually derived rules, where the latter target very sparse emotion categories. The first approach considers the task as single label multi-class classification, where an SVM and a CRF classifier are trained to recognise fifteen different categories and their results are combined. Our second approach trains individual binary classifiers (SVM and CRF) for each of the fifteen sentence categories and returns the union of the classifiers as the final result. Finally, our third approach is a combination of binary and multi-class classifiers (SVM and CRF) trained on different subsets of the training data. We considered a number of different feature configurations. All three systems were tested on 300 unseen messages. Our second system had the best performance of the three, yielding an F1 score of 45.6% and a Precision of 60.1% whereas our best Recall (43.6%) was obtained using the third system.

Keywords: emotion classification, hybrid, suicide, sentence classification

Biomedical Informatics Insights 2012:5 (Suppl. 1) 175–184

doi: 10.4137/BII.S8967

This article is available from http://www.la-press.com.

Introduction

Suicide is one of the leading causes of death worldwide and presents an increasingly serious public health problem in developed countries.[1] Doctors and other caregivers stand in the front line in the battle to prevent tragedy, facing the urgent need to determine from scarce information the risk of a successful attempt, such that preventative measures can be undertaken. The emotional state of the patient is highly relevant in this task, since depression and other disorders of emotional functioning are known to substantially raise the risk of suicide. Tools which are able to automatically process emotional state in textual resources will be invaluable in the medical fight to intercept such states.[2,3] The 2011 i2b2 Medical NLP Challenge presents participants with the task of classifying emotions at the sentence level as they appear in a corpus of suicide notes collected during medical research. The training data consists of messages manually annotated with several different categories of emotion as well as two non-emotion categories: information and instructions. The annotation was performed by relatives of the victims.

There are multiple challenges in automatically assigning one of the 15 possible categories to a sentence in the notes. The annotation scheme is mixed, covering different categories of emotion as well as non-emotion categories; the guidelines, instructions and criteria followed by annotators have not been disclosed; there is apparent ambiguity between instructions and information; sentences are often annotated with more than one category, rendering their processing a multi-label classification task; and many of the categories are very sparse.

Our approach was motivated by these considerations. We first examined the distribution of the fifteen categories in the training data. The majority of annotated sentences (74%, 37.45% of total) are covered by just four categories, namely 'instructions', 'hopelessness', 'love' and 'information', while another three categories ('guilt', 'blame' and 'thankfulness') account for another (16%). So a total of 90% of the annotated data are described by only seven of the fifteen categories while 49% of all sentences received no annotation. A clear dilemma was whether to aim to maximise system performance by focusing on the four to seven major categories, or to try and address all emotion categories. Since this was a new task involving an important and controversial topic, we saw it as an opportunity to explore the annotation scheme and make observations about the types of features suited to the recognition of different types of emotions. We decided to follow a hybrid approach which would consider state of the art supervised machine learning as well as manually created rules to cater for the sparse emotion categories. We believe that our approach provides a good insight to the various emotion categories present in suicide notes and can lay the foundations for future applications which would make use of such emotion recognition.

Related Work

Recent work has described the use of text mining approaches to differentiate genuine suicide notes from simulated ones, finding that machine learning approaches were able to make the discrimination better than humans in some cases.[3] Both content-based and structural features were used in the classification. The authors observe that human discriminators focused primarily on content-based features (such as the concepts annotated from the ontology) in making the discrimination, while the machine learning approach obtained the highest information gain from the structural approaches in this task. However, as our task here is emotion classification rather than genuine note detection, we emphasised content-related features more highly in our selection.

The distinction between content and structural features is emphasised in the work of Shapero[4] which describes an extensive investigation into the language used in suicide notes, and reports on the features found more commonly in genuine notes than simulated notes. Genuine notes are found to include affection, the future tense, references to family members, pronouns, names, negatives, intensifiers and maximum quantity terms. Some structural features such as the presence of dates or the identity of the author were found to be more common in genuine notes.

With a promising approach to early intervention using the increasing online presence of particularly teenagers and young adults on Web 2.0, Huang et al describe a simple approach with dictionary-based keyword detection to automatically detect suicide risk and flag depression from blog posts and posts to popular social networks.[5,6] This compares closely to the manual rule approach which we have adopted for the several least populous categories in our training data, which also relies on encoded keywords and phrases. However, Huang et al focus on one emotional category only, which simplifies their task, as they do not directly face disambiguation problems.

Another study aimed to automatically distinguish suicide notes from arbitrary newsgroup articles, as part of a broader effort to develop tools which can distinguish suicidal communication from general conversation.[2] This research used words and grammatical features which were automatically discovered in the corpus, then clustering features across the suicide notes and the newsgroup articles, showing clear divisibility in semantic space. Importantly, the clustering results also showed subcategories within suicide notes for those which are emotional and those which are unemotional, providing some incentive for studying the emotional expressions in suicide notes.

Automatically classifying emotions in suicide notes is a special case of emotion detection in text, a task which has applications in human-computer interaction and sentiment analysis for marketing research.[7,8] While such work is closely related to this project, it differs in the nature of the classification to be performed and the text to be classified. In some cases, only *positive emotion* and *negative emotion* are used as grouping classes, and in others, only a small set of basic categories of emotion are used to minimise semantic overlap, eg, anger, joy, disgust, sadness and fear.[8] Interestingly, in[8] they find that word stemming actually reduces the effectiveness of classifiers as in some cases the emotional meaning of the word is altered; also, it was noted that word tense can be important. But such emotion classification projects benefit from much larger training corpora, and another challenge which the suicide note medium presents that it not usually faced by these other emotion classification tasks is the low structural quality of the language.

Methods

We first considered single label multi-class classification, where sentences with multiple categories appeared in the training data as multiple copies, a reasonable first step as annotated sentences in the training data have 1.16 labels on average. We employed both JRip and SMO in weka[9] and also LibSVM and CRFSuite, for which we obtained higher performance (average f1 of 0.4425). We also trained binary classifiers for each emotion and considered their union in order to label the data with emotion categories. The latter approach permitted the assignment of multiple labels and also is better suited to imbalanced data. Indeed, we obtained our highest performance in this way (average f1 of 0.46). Another approach we considered was training both binary classifiers and multi-class classifiers on different subsets of the data and combining them to obtain class assignments. While this approach is promising both in terms of multiple label assignment and increased recall, it generated many false positives and achieved an average f1 measure of 0.3977.

Data

Our training data consist of 600 suicide messages of varying length and readability quality, ranging from a single sentence to 177, with over 80% of messages containing fewer than 10 sentences and the average message length being 4 sentences. The messages have been labelled at the sentence level with one or more of 15 categories while a large percentage (49.38%) have received no annotation. Label cardinality is 0.54 overall and 1.16 for annotated sentences, making multiple annotations rare. As humans reading the texts we found that the distinction between information, instruction and sentences without any annotation was unclear. We pre-processed such sentences to facilitate feature extraction by replacing all names, times and places with the words NAME, TIME and ADDRESS respectively.

Features

We employed a number of sentence based features as input to machine learning classifiers, many of which have been used in other types of text classification. Single words and bigrams extracted from a sentence are the default features and thus our baseline considers only ngrams. Other work on sentence based classification, such as argumentative zoning,[10] has shown that ngram-based systems are hard to beat. We also considered grammatical features such as verbs, the tense and voice of a verb (both of which have been shown to be significant in the classification of scientific texts[11,12] as well as subjects, direct objects and indirect objects of verbs and grammatical triples. The latter consist of the type of the dependency relation (eg, subj, obj, iobj), the head word and the dependent word. We anticipated that the latter relations would help detect patterns which could distinguish between self-directed emotions and emotions geared towards others. To obtain the parts of speech and grammatical relations we used C&C tools.[13] We also introduced a negation feature, which denotes the presence or absence of a negative expression in a sentence. Negative expressions were annotated automatically using.[14] Negation is particularly relevant to the detection of certain emotions which are often followed by a negative expression. Indeed, overall performance increased with the addition of the negation feature. We also implemented length as a feature of the sentence, as different categories tend to have different average word length. Finally, we also took into account two global sentence features, namely the location of a sentence within a message, split into five equal segments, and the category of the previous sentence (history). The history feature was implemented when we used SVM as a classifier, to model the sequence of categories within a message. This feature was abandoned as soon as we established that SVMs performed better without it (up to 5% higher f-measure), which was due to propagation of error from preceding erroneous predictions.

Classifiers

We used LibSVM[1], coded in C++ since with the same features it performed better than SMO in weka. Our experiments were conducted using a linear kernel, known to perform well in document classification. We used the default values for the C and ∈ parameters and concentrated on the input features. The recent BioNLP challenge,[15] which addresses a series of tasks ranging from event extraction to coreference resolution has shown the importance of input features as performance of the same classifier can dramatically according to the features.

A drawback in using SVMs is that one cannot easily model the sequence of categories in a message without introducing errors as was the case with the history feature. While we could not be certain that suicide messages are structured, as scientific texts are, one of our hypotheses is that certain emotion categories tend to follow others or tend to cluster together. For this reason we employed Conditional Random Fields (CRF), which have been shown to give good results in sequence labeling of abstracts.[16] We used CRFSuite,[2] an algorithm for linear-chain, First Order CRFs, optimised for speed and implemented in C. Stochastic Gradient Descent was employed for parameter estimation.

Both LibSVM and CRFSuite were used in a number of different configurations, both for single label multi-class classification and multi-label classification (Fig. 1).

Multi-class annotation

We trained LibSVM and CRFSuite models independently of each other but on the same training data single label multi-class classification. We combined the results of the two classifiers so that in cases of disagreement we chose the category that had received the highest probability, according to classifier output.

Binary classifiers

As instances in the training data contain multiple annotations per sentence, we trained individual binary classifiers for each of the fifteen categories present in the training data, for both LibSVM and CRFSuite and took the union of the classifiers. This allowed a sentence to receive more than one category if more than one binary classifier made a class assignment. Category assignments from LibSVM and CRFSuite were considered

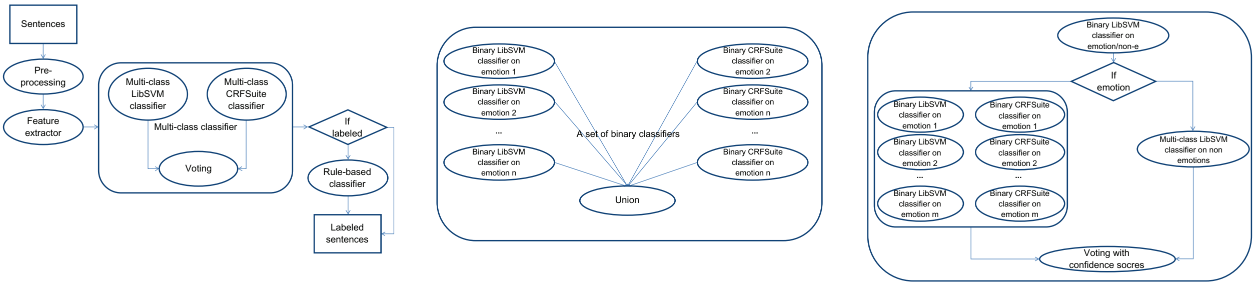
Sentences
Pre-processing
Feature extractor
Multi-class LibSVM classifier
Multi-class CRFSuite classifier
Multi-class classifier
Voting
If labeled
Rule-based classifier
Labeled sentences
Binary LibSVM classifier on emotion 1
Binary LibSVM classifier on emotion 2
...
Binary LibSVM classifier on emotion n
A set of binary classifiers
Binary CRFSuite classifier on emotion 2
Binary CRFSuite classifier on emotion n
...
Binary CRFSuite classifier on emotion n
Union
Binary LibSVM classifier on emotion/non-e
If emotion
Binary LibSVM classifier on emotion 1
Binary CRFSuite classifier on emotion 1
Binary LibSVM classifier on emotion 2
Binary CRFSuite classifier on emotion 2
...
...
Binary LibSVM classifier on emotion m
Binary CRFSuite classifier on emotion m
Multi-class LibSVM classifier on non emotions
Voting with confidence socres

Figure 1. Multi-class single label/Union of binary classifiers, multi-label/Combination of binary and multiclass classifiers, multi-label.

as separate if they involved different categories. This approach yielded the best performance.

Combination of binary and multi-class classifiers

We also implemented a variant of the binary classifiers, which combines binary classifiers trained on subsets of the data with a multi-class classifier trained on another subset. We call this approach 'Hybrid Binary' and is a variant on hierarchical classification. We first trained a single binary classifier on the training data, to distinguish between emotion and non-emotion sentences. We then trained individual emotion classifiers, using only the subset of the data pertaining to emotions, and a multi-class classifier on the non-emotion data, which determined whether a sentence should receive information, instructions, information-instructions or remain unannotated. A sentence was assigned the union of the output of the emotion classifiers. In cases where the binary non-emotion classifier fires, without the binary emotion classifier firing as well, the sentence is also assigned the category determined by the non-emotion multi-class classifier. This more sophisticated approach was intended to boost emotion recognition and indeed resulted in the best recall we achieved (43.6%). The drawback was that it also generated many false positives.

Manual rules

As several of the emotion categories were extremely sparsely annotated relative to the corpus as a whole, the classifiers were unable to return meaningful results for these categories. We thus decided to complement the machine learning approach with a dictionary of manual recognition rules for high precision, low recall emotion recognition. The rules were proposed by manual inspection of the relevant annotated sentences from the training data together with the examination of synonym sets from WordNet-Affect.[17] Each rule was then validated by testing against the corpus as a whole, and rules which were "noisy" were discarded.

We found that the type of language used in emotional sentences varied strongly depending on the emotion category. For example, the language used in the 'love' and 'forgiveness' categories was quite homogeneous (almost every sentence containing the words 'love' and 'forgive' respectively), while the language used in other categories such as 'anger' was extremely heterogeneous and metaphorical. There were also large overlaps in the language used between different categories which were close in semantic meaning, such as between 'anger' and 'blame', and between 'happiness peacefulness' and 'hopefulness'.

We used 48 manual rules for the 8 sparsest categories, distributed across the categories: anger (12), sorrow (5), hopefulness (9), happiness_peacefulness (4), fear (6), pride (6), abuse (4), forgiveness (2). These manual rules, developed in Perl regular expressions, were only applied to those sentences not labelled by automatic classifiers, meaning that the number of sentences applied to the rules varies over automatic classifiers.

Results and Discussion

Table 1 shows the outcome of three different approaches each with three different feature configurations (all features, ngrams, and all features without negation) on the training data. We obtained the best F-measure from the 'Binary All' approach, which took the union of individual binary SVM and CRF classifiers trained with all features. From the multi-class classifier we can see that CRF performed better than

Table 1. Results on training data.

	Feature	SVM	CRF	SVM+CRF	SVM+CRF+man
	All	0.3957	0.4347	0.4415	0.438
Multi class	Ngram	0.3606	0.4416	0.4437	0.44
	No-neg	0.3665	0.4332	0.4349	0.4318
	All			0.464	0.461
Binary all	Ngram			0.46	0.457
	No-neg			0.46	0.457
	All			0.39	0.389
Hybrid binary	Ngram			0.384	0.384
	No-neg			0.3875	0.3865

Table 2. Results on test data.

Precision	Recall	F-measure	Annotations
0.56366	0.35849	0.43825	809
0.60077	0.36792	0.45636	779
0.36251	0.43632	0.39600	1,531

SVM, which suggests that the sequence of sentences and categories does play a role in emotion detection in the suicide notes.

We also analysed the individual performance of the 'Binary All' approach for each category when all features were used and how it was influenced in the presence of a single feature (each of GR (grammatical triple), Subject, Verb, and Negation). Table 3 shows the result of the 'Binary All' classification for each category and Table 4 the same result combined with manual rules. For all major categories (instructions-thankfulness) the best results were obtained for the combination of all features but the difference between ngrams and all features is not very big, ranging from 1%–4%. For rare categories (anger to forgiveness) the combination with manual rules outperforms the machine learning classifier-only approach. However, this combination with manual rules reduces overall performance by increasing FP rather than reducing FN in the case of all, ngram and unigram features. In the case of the GR, Subject, Verb and Negation features, where using only ML classifiers produced no results, performance for rare categories increased.

We believe that we could improve our results by finding better ways of combining classifiers, perhaps through stacking or joint inference, techniques which achieved the highest results in BioNLP 2011.[15] Judging from manual rule-only results for the rare categories, we believe that a hybrid system which combines machine learning predictions for the major categories with manual rule-only results for the rare categories could boost recognition performance.

For the test data submission, we first chose the best feature model from each category of classifiers (ie, multi-class, binary and hybrid binary) and applied it to the test data. We separately applied the manual rules to the test data. We then combined the output of each model with the manual rules so that the manual rules applied only to sentences where the classifiers had made no predictions. The results we obtained from the scoring website are in Table 2.

Data inconsistency

We observed several inconsistencies in the training data, a factor which we believe led to decreased performance in the resulting machine learning and rule-based approaches. Structurally, it is noticeable that the "sentence-level" annotation often transcends sentences. For example, *The cards were just stacked against me. Honey Get insurance on furniture soon as you can* is

Table 3. Binary classification result for each class.

	All			Ngram			Unigram		
	P	R	F	P	R	F	P	R	F
Instructions	0.6329	0.5334	0.5789	0.6437	0.5545	0.5957	0.5766	0.5681	0.5723
Hopelessness	0.547	0.372	0.4429	0.5064	0.3744	0.4305	0.4534	0.4265	0.4396
Love	0.6483	0.5312	0.584	0.6303	0.5208	0.5703	0.5679	0.566	0.567
Information	0.5302	0.2862	0.3718	0.4641	0.2572	0.331	0.3502	0.3007	0.3236
Guilt	0.4742	0.2447	0.3228	0.4103	0.2553	0.3148	0.2764	0.2926	0.2842
Blame	0.3158	0.0606	0.1017	0.25	0.0707	0.1102	0.1481	0.1212	0.1333
Thankfulness	0.7544	0.4624	0.5733	0.7097	0.4731	0.5677	0.589	0.4624	0.5181
Anger	0	0	0	0	0	0	0.0263	0.0159	0.0198
Sorrow	0	0	0	0	0	0	0.0625	0.0612	0.0619
Hopefulness	0	0	0	0	0	0	0.069	0.0455	0.0548
Happiness_ peacefulness	0	0	0	0	0	0	0.0909	0.0435	0.0588
Fear	0	0	0	0	0	0	0	0	0
Pride	0	0	0	0	0	0	0	0	0
Abuse	0	0	0	0	0	0	0	0	0
Forgiveness	0	0	0	1	0.1667	0.2857	0.5	0.1667	0.25
Overall	0.5957	0.3806	0.4645	0.5708	0.3856	0.4603	0.4545	0.4172	0.435

included as one sentence, while clearly should have been separated into two. This "sentence" received a double annotation (hopelessness and instructions), which would have been separate annotations if the sentences had been properly separated. The converse also appears.

There were also inconsistencies in the annotation of categories to the sentences. We observed significant ambiguities between the (most voluminous) non-emotional categories 'information' and 'instructions'. Some sentences were annotated with both categories, such as *John my books are up under the cash register*. This sentence does contain information, but to the casual reader, it is not obvious what is instructional in this sentence. Conversely, *In case anything happens please call my attorney—John Johnson—9999 3333 Burnet Ave* is annotated with both but appears solely instructional. Yet other sentences annotated with only one of the two appeared to us to have had the incorrect choice of category applied, while some sentences appeared to contain information or instructions but were un-annotated.

Within emotional categories, several sentences which are very similar were inconsistently annotated, for example the phrase phrase *God forgive me* was annotated as 'guilt' (sometimes combined with 'hopelessness') in several sentences including *My God forgive me for all of my mistakes*, but makes one appearance with no annotation and one (separate note) as 'instructions' for *May God forgive me. Take care of them*, and another as 'hopefulness' in *May God forgive me, and I pray that I mite be with my wife for ever when we leave this earth*.

Annotation Guidelines

Sentence-level annotation of classification categories in free text is an intrinsically difficult task, and quality of annotations need to be ensured by interannotator agreement values. In an ideal corpus for machine learning, consistency in annotation is required. Emotion language is deeply ambiguous and open to diverse interpretations. Furthermore, a sentence might express that the writer was feeling a certain way when they wrote the text, although this is not in itself explicit in the text. Many of the sentences which were labelled with anger were labelled as such because the tone of the sentence seemed angry, not because anger was explicitly mentioned: the word "angry" does not appear even once in the corpus of 69 annotated sentences. On the other hand, the statement *I was always afraid to ...* directly expresses fear, although that might not be what the author was experiencing at the time that they wrote the sentence. To achieve consistency, annotation guidelines should clarify intended scenarios for different categories. A relevant project in this area is the *emotion ontology* which is being developed to facilitate annotation of emotions in text.[18] Such an ontology is not an annotation scheme in itself, but provides *definitions* which can be used for

GR			Subject			Verb			Negation		
P	R	F	P	R	F	P	R	F	P	R	F
0.4961	0.1584	0.2402	0.5562	0.1163	0.1924	0.5045	0.1399	0.2190	0	0	0
0.2908	0.0972	0.1456	0.4493	0.0735	0.1263	0.2778	0.0474	0.081	0	0	0
0.4537	0.1701	0.2475	0.1667	0.0035	0.0068	0.4762	0.0347	0.0647	0	0	0
0.4444	0.087	0.1455	0.3659	0.0544	0.0946	0.2727	0.0326	0.0583	0	0	0
0.0899	0.0426	0.0578	0.25	0.0213	0.0392	0.1458	0.0372	0.0593	0	0	0
0.1053	0.0202	0.0339	0	0	0	0.1250	0.0202	0.0348	0	0	0
0.2069	0.0645	0.0984	0.3636	0.086	0.1391	0.2424	0.086	0.127	0	0	0
0	0	0	0.2	0.0318	0.0548	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0.0952	0.0909	0.0930	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0.3417	0.1073	0.1633	0.4247	0.0645	0.112	0.3346	0.0711	0.1173	0	0	0

Table 4. Binary classification result combined with manual rules for each class.

	All			Ngram			Unigram			GR		
	P	R	F	P	R	F	P	R	F	P	R	F
Instructions	0.6329	0.5334	0.5789	0.6437	0.5545	0.5957	0.5766	0.5681	0.5723	0.5	0.16	0.24
Hopelessness	0.547	0.372	0.4429	0.5064	0.3744	0.4305	0.4534	0.4265	0.4396	0.29	0.1	0.15
Love	0.6483	0.5312	0.584	0.6303	0.5208	0.5703	0.5679	0.566	0.567	0.45	0.17	0.25
Information	0.5302	0.2862	0.3718	0.4641	0.2572	0.331	0.3502	0.3007	0.3236	0.44	0.09	0.15
Guilt	0.4742	0.2447	0.3228	0.4103	0.2553	0.3148	0.2764	0.2926	0.2842	0.09	0.04	0.06
Blame	0.3158	0.0606	0.1017	0.25	0.0707	0.1102	0.1481	0.1212	0.1333	0.11	0.02	0.03
Thankfulness	0.7544	0.4624	0.5733	0.7097	0.4731	0.5677	0.589	0.4624	0.5181	0.21	0.06	0.1
Anger	0.1667	0.0318	0.0533	0.1333	0.0318	0.0513	0.0652	0.0476	0.0551	0.06	0.03	0.04
Sorrow	0.1429	0.1429	0.1429	0.1228	0.1429	0.1321	0.0899	0.1633	0.1159	0.2	0.24	0.22
Hopefulness	0.1304	0.0682	0.0896	0.125	0.0682	0.0882	0.087	0.0909	0.0889	0.09	0.07	0.08
Happiness_peacefulness	0.0769	0.0435	0.0556	0.0833	0.0435	0.0571	0.087	0.087	0.087	0.06	0.04	0.05
Fear	0.2	0.1818	0.1905	0.2	0.1818	0.1905	0.1304	0.1364	0.1333	0.14	0.18	0.16
Pride	0.6667	0.1333	0.2222	1	0.2	0.3333	1	0.0667	0.125	0.4	0.13	0.2
Abuse	0.5	0.375	0.4286	0.25	0.125	0.1667	0.5	0.25	0.3333	0.6	0.38	0.46
Forgiveness	0.2222	0.3333	0.2667	0.3	0.5	0.375	0.1667	0.1667	0.1667	0	0	0
Overall	0.5653	0.3906	0.462	0.5425	0.3952	0.4573	0.4415	0.4239	0.4325	0.32	0.12	0.17

definitive disambiguation between similar categories, such as 'blame' and 'anger'. An ontology specifically for suicide note annotation is proposed and used in.[3] It includes some of the same emotion categories used for annotation in this challenge, although it is more extensive, including categories such as 'self aggression' and 'helplessness'. However, it is not clear in[3] whether the ontology terms are accompanied by disambiguating definitions. Annotation guidelines should also clarify the objective of the natural language processing. On the one hand, if the purpose is to obtain the best performance from an NLP system for emotion identification in itself, the emotions with low prevalence can be regarded as essentially irrelevant. On the other hand, if the objective of the task is to study emotions in the context of suicide, even low-prevalent emotions may bear scientific interest.

It is of general interest that the principal emotion found in this suicide note corpus is hopelessness. This can be compared to the result of,[3] who find that the most relevant emotion categories for detecting genuine notes are: giving things away, hopeless, regret and sorrow. However, detecting emotions such as hopelessness in human text is inherently plagued by the flexibility of words such as "hope" and "wish". Both[3] and[4] find a surprising role for structural features in real suicide notes—which are not obviously emotional in nature. A parallel in the current task is that the highest prevalence is instructions in the notes. It would be surprising, however, if the same features worked equally well for such non-emotional content as for detecting the emotional sentences.

Conclusions and Future Work

In this paper, we described three different approaches to detect emotions from sentences, motivated by the sparse and imbalanced data as well as the complex annotation scheme. In each approach, we explored various feature representations from simple unigrams to rich grammatical information, and we also tested the use of negation which can change the meaning of sentences.To improve the performance for rare categories, we wrote manual rules and combined them with ML-based classification results. We found some interesting results which encourage further investigating the use of manual rules for rare categories.

As future work, we plan to explore various methods of integrating different machine learning classifiers for emotion recognition, using techniques such as stacking and joint inference. We would also like to experiment with different techniques for combining manual rules with automatic classifiers more systematically. We will test the use of meta-classifiers using results of individual

Subject			Verb			Negation			Manual Only		
P	R	F	P	R	F	P	R	F	P	R	F
0.56	0.12	0.19	0.5	0.14	0.22	0	0	0	0	0	0
0.45	0.07	0.13	0.28	0.05	0.08	0	0	0	0	0	0
0.17	0	0.01	0.48	0.03	0.06	0	0	0	0	0	0
0.37	0.05	0.09	0.27	0.03	0.06	0	0	0	0	0	0
0.25	0.02	0.04	0.15	0.04	0.06	0	0	0	0	0	0
0	0	0	0.13	0.02	0.03	0	0	0	0	0	0
0.36	0.09	0.14	0.24	0.09	0.13	0	0	0	0	0	0
0.21	0.08	0.11	0.08	0.03	0.04	0.07	0.05	0.06	0.1875	0.0476	0.0759
0.17	0.27	0.21	0.14	0.22	0.18	0.18	0.29	0.22	0.1772	0.2857	0.2186
0.11	0.09	0.1	0.14	0.11	0.13	0.18	0.14	0.16	0.1818	0.1364	0.1558
0.04	0.04	0.04	0.06	0.04	0.05	0.05	0.04	0.05	0.0526	0.0435	0.0476
0.22	0.27	0.24	0.16	0.32	0.22	0.23	0.27	0.25	0.2308	0.2727	0.25
0.6	0.2	0.3	0.43	0.2	0.27	1	0.2	0.33	1	0.2	0.3333
0.33	0.13	0.18	0.5	0.25	0.33	0.6	0.38	0.46	0.6	0.375	0.4615
0.15	0.33	0.21	0.1	0.17	0.13	0.25	0.5	0.33	0.25	0.5	0.3333
0.35	0.08	0.13	0.3	0.08	0.13	0.18	0.02	0.03	0.2021	0.0162	0.03

classifiers as features of ML-based classifiers. It will also be interesting to test if more elaborate feature selection could help improve the results.

We believe that our work can provide insight into the recognition of various emotion categories present in suicide notes and can benefit applications relating to emotion recognition in blogs and other personal statements.

Disclosures

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

References

1. National institute of mental health—suicide in US: Statistics and prevention. URL http://www.nimh. nih.gov/health/publications/suicide-in-the-us-statistics-and-prevention/ index.shtml.
2. Matykiewicz P, Duch W, Pestian J. Clustering semantic spaces of suicide notes and newsgroups articles. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, Stroudsburg, PA, USA; 2009:179–84. Association for Computational Linguistics. ISBN 978-1-932432-30-5.
3. Pestian John, Nasrallah Henry, Matykiewicz Pawel, Bennett Aurora, Leenaars Antoon. Suicide note classifi-cation using natural language processing: A content analysis. *Biomed Inform Insights*. 2010:19–28.
4. Shapero Jess Jann. *The language of suicide notes*. PhD thesis, University of Birmingham, July 2011. URL http://etheses.bham.ac.uk/1525/. Appendix C is not available online, nor is it available for consultation in the Library.
5. Huang Yen-Pei, Goh Tiong, Liew Chern Li. Hunting suicide notes in web 2.0—preliminary findings. In: *Proceedings of the Ninth IEEE International Symposium on Multimedia Workshops, 2007, Beijing ISMW* '07; 2007: 517–21.
6. Goh Tiong-Thye, Huang Yen-Pei. Monitoring youth depression risk in web 2.0. *VINE*. 2009;39:192–202.
7. Pang Bo, Lee Lillian. Opinion mining and sentiment analysis. *Found Trends Inf Retr*. Jan 2008;2:1–135. ISSN 1554-0669. doi: 10.1561/1500000011.
8. Danisman Taner, Alpkocak Adil. Feeler: Emotion Classification of Text Using Vector Space Model. In *AISB* 2008 *Convention, Communication, Interaction and Social Intelligence*, volume 2, Aberdeen, UK, Apr 2008.
9. Hall Mark, Frank Eibe, Holmes Geoffrey, Pfahringer Bernhard, Reutemann Peter, Witten Ian H. The WEKA data mining software: an update. *SIGKDD Explor Newsl*. Nov 2009;11:10–8. ISSN 1931-0145. doi: http://doi.acm.org/10.1145/1656274.1656278.
10. Merity Stephen, Murphy Tara, Curran James R. Accurate argumentative zoning with maximum entropy models. In: *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, NLPIR4DL '09, Morristown, NJ, USA; 2009:19–26. Association for Computational Linguistics. ISBN 978-1-932432-58-9.
11. Teufel Simone, Moens Marc. Summarizing scientific articles: experiments with relevance and rhetorical status. *Comput Linguist*. Dec 2002;28: 409–45. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/08912010276 2671936.
12. Guo Y, Korhonen A, Liakata M, Silins I, LiSun L, Stenius U. A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. *BMC Bioinformatics*. 2011.

13. Curran James, Clark Stephen, Bos Johan. Linguistically motivated large-scale nlp with c&c and boxer. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic; Jun 2007: 33–6. Association for Computational Linguistics.
14. Morante Roser, Van Asch Vincent, Daelemans Walter. Memory-based resolution of in-sentence scopes of hedge cues. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, CoNLL '10: Shared Task, Stroudsburg, PA, USA; 2010:40–7. Association for Computational Linguistics. ISBN 978-1-932432-84-8.
15. Kim Jin-Dong, Pyysalo Sampo, Ohta Tomoko, Bossy Robert, Nguyen Ngan, Tsujii Jun'ichi. Overview of bionlp shared task 2011. In: *Proceedings of BioNLP Shared Task* 2011 *Workshop*, Portland, Oregon, USA; Jun 2011:1–6. Association for Computational Linguistics.
16. Hirohata K, Okazaki N, Ananiadou S, Ishizuka M. Identifying sections in scientific abstracts using conditional random fields. In: *Proc of the IJCNLP.* 2008.
17. Valitutti RO. Wordnet-affect: an affective extension of wordnet. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation*. 2004;1083–1086.
18. Hastings Janna, Ceusters Werner, Smith Barry, Mulligan Kevin. Dispositions and processes in the Emotion Ontology. In: *Proceedings of the International Conference on Biomedical Ontology (ICBO 2011), Buffalo, U S A*. 2011.